01｜打好基础：到底什么是机器学习？

2021-08-30 黄佳

《零基础实战机器学习》 课程介绍 >

讲述：黄佳
时长 20:36 大小 18.87M

你好，我是黄佳。

欢迎来到《零基础实战机器学习》！在开篇词里面，我说过学机器学习的秘诀是“做中学”。不过呢，即使上来就想要上手做，最最最基础的知识你还是得懂点。

说是基础知识，其实你也不用怕。我们这一讲里要讲的知识点不多、也不难，主要是想讲讲什么是机器学习和一些最常见的概念。你不知道这些的话，肯定是没法开始实战的。

除此之外，我想讲讲机器学习的分类。我讲这些，是想让你在实战之前对机器学习的大心中有数，方便你一边实战，一边构建自己的知识图谱。同时，也可以让你提前了解下我们的实战项目，对我们将来要闯的关卡有个预期。而至于更多的基础知识，我会带你边做项目边学习。

在开始之前，我们先给这一讲设一个小目标，就是你在学完之后，当别人问你什么是机器学习的时候，你要能够给他解释明白。真能把这个说清楚了，你这节课也就没白学了。

什么是机器学习

这个问题其实不好回答，因为机器学习涵盖的内容太多了。机器学习之父 Arthur Samuel 对机器学习的定义是：在没有明确设置的情况下，使计算机具有学习能力的研究领域。国际机器学习大会的创始人之一 Tom Mitchell 对机器学习的定义是：计算机程序从经验 E 中学习，解决某一任务 T，进行某一性能度量 P，通过 P 测定在 T 上的表现因经验 E 而提高。

这两个定义你看了之后可能瞬间就懵了，没关系，这里我用“人话”来和你解释一下。

现在，请你想象这样一个场景：你周日约了小李、老王打牌，小李先来了，老王没来。你想打电话叫老王过来。小李说：“你别打电话啦，昨天老王喜欢的球队皇马输球了，他的项目在上个礼拜也没成功上线，再加上他儿子期末考试不及格，他肯定没心情来。”

这种情况下，你觉得老王会来吗？

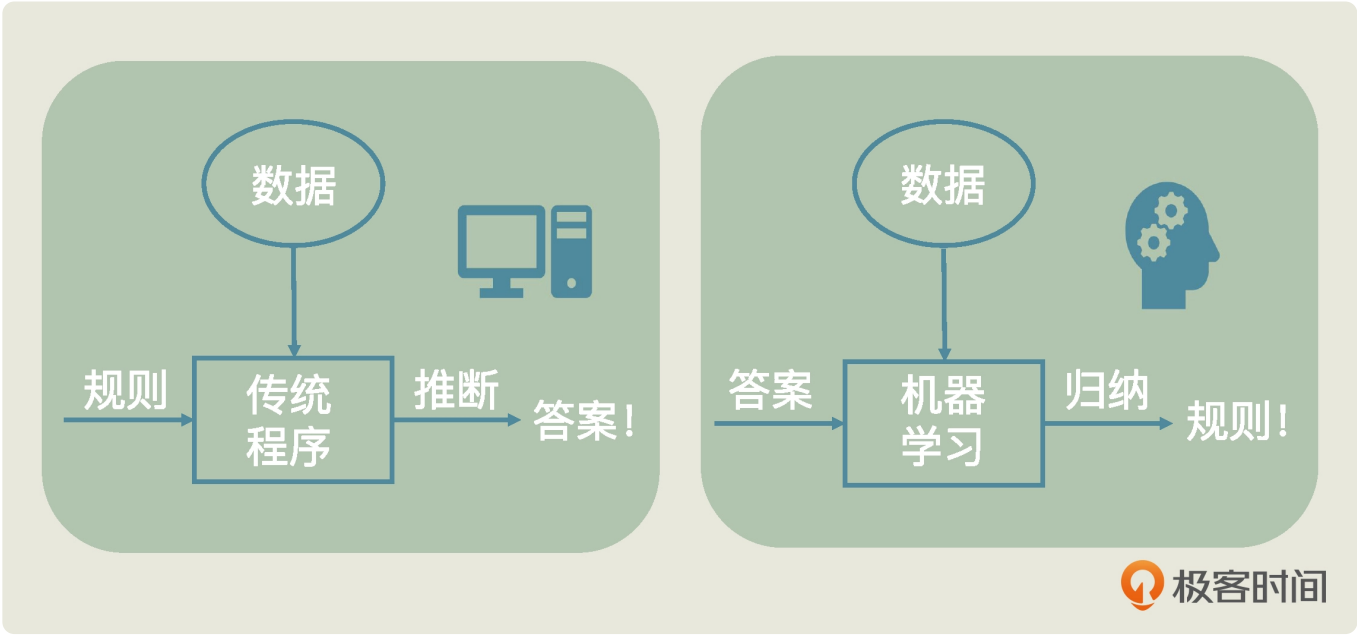
一般情况下，我们都会觉得老王大概率不会来了。不过，你有想过我们是怎么得出这个结论的吗？

实际上我们运用了“推理”。我们人类的大脑做这样的推理似乎是自然而然的事儿。但是，对于计算机来说，如果它也像小李那样有老王的“历史数据”，知道他看皇马，知道他的项目情况，知道他儿子的成绩，那计算机能推出这个结论吗？对于长期以来只能按照人类预设规则解决问题的机器来说，这可并不是件容易的事。

而机器学习的厉害之处就在于，它能**利用计算机的运算能力，从大量的数据中发现一个“函数”或“模型”，并通过它来模拟现实世界事物间的关系，从而实现预测、判断等目的**。这个过程的关键是建立一个合适的模型，并能主动地根据这个模型进行“推理”，而这个建模的过程就是机器的“学习”过程。

那么机器学习和我们传统的程序有什么区别呢？实际上，传统程序是程序员把已知的规则定义好后输入给机器的，而机器学习则从已知数据中，通过不断试错、自我优化、自身总

结，归纳出规则来。下面这张图，直观地阐述了机器学习和传统程序的区别，你可以看一下。

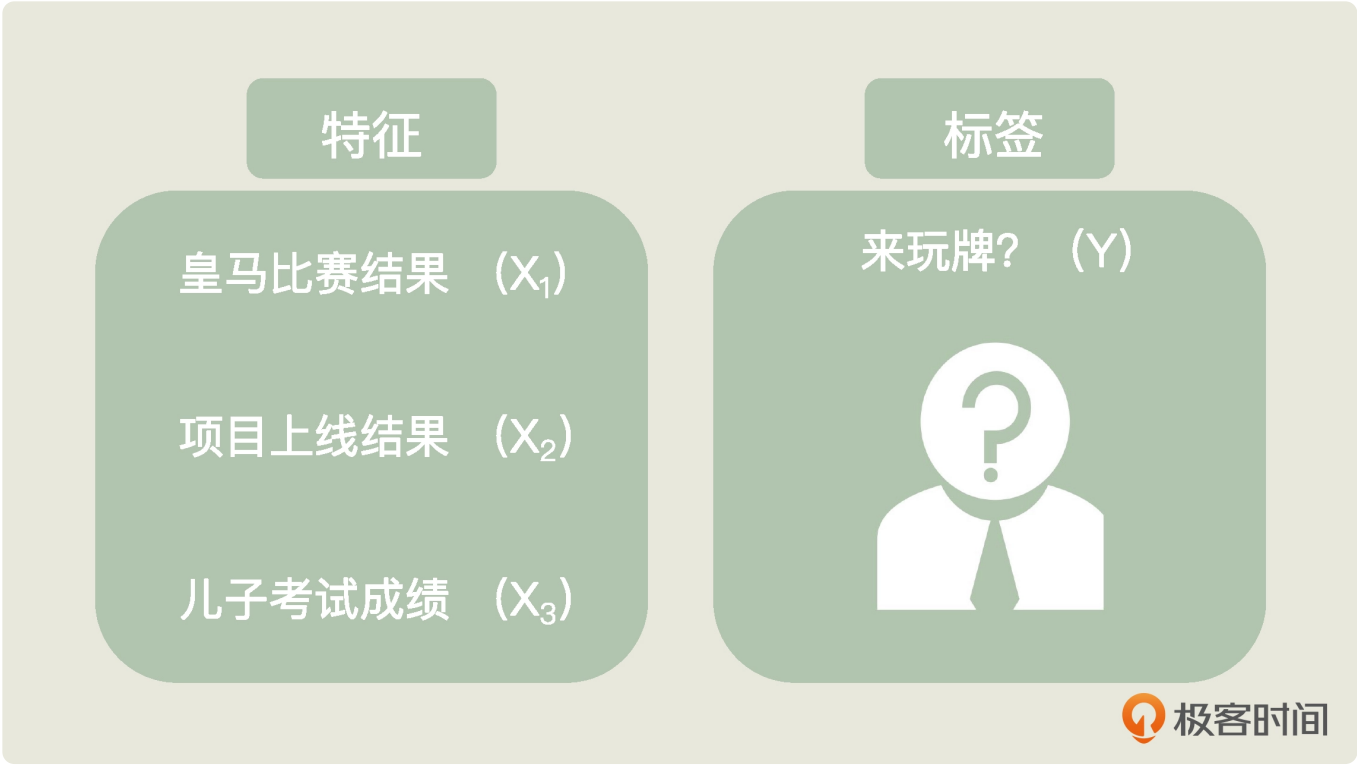


这张图中展示了机器学习的本质特征，就是**从数据中发现规则**。

虽然说我们希望机器最终能主动地去预测，但在此之前，机器具体选什么模型、如何训练、怎么调参，我们人类还是要在这个过程中给机器很多指导的，这就是我这门课要教你的。

不过，机器到底怎样建立新的模型呢？我们继续以刚才的场景为例，不过，现在我想请你从函数的角度想一想刚才发生了什么？

要预测老王的状况，我们就需要建立一个“预测老王会不会来”的函数，而“皇马输赢”、“项目情况”、“儿子成绩”都是输入到这个函数的自变量，我们设为 x1, x2, x3。这些自变量每一个发生变化，都会影响到函数的结果，也就是因变量 y。



在机器学习中，这些自变量，就叫做特征（feature），因变量 y 叫做标签（label）。而一批历史特征和一批历史标签的集合，就是机器学习的数据集。

理解了这些，我们就可以更加“精准”地定义机器是怎么“学习”的了，就是在已知数据集的基础上，通过反复的计算，选择最贴切的函数去描述数据集中自变量 x1, x2, x3, ..., xn 和因变量 y 之间的因果关系。这个过程，就叫做机器学习的训练，也叫拟合。

基于这一点，我们可以说：**传统程序是程序员来定义函数，而在机器学习中是机器训练出函数。**

最初用来训练的数据集，就是训练数据集（training dataset）。当机器通过训练找到了一个函数，我们还需要验证和评估，也就是说，这时候我们要给机器另一批同类数据特征，看机器能不能用这个函数推出这批数据的标签。这一过程就是在验证模型是否能够被推广、泛化，而此时我们用到的数据集，就叫验证数据集（validation dataset）。

简单来说，在验证、评估的过程里，我们就是要验证这个函数到底好不好。如果这个函数通过了评估，那就可以在测试数据集（test dataset）上做最终的测试；如果通过不了，就需要继续找新的模型。

讲到这里，我想你已经对机器学习有了一定的了解。不知道你有没有发现，标签似乎对于机器学习模型有很重要的指导性意义，因为机器必须根据已有的数据来找到特征和标签之

间的关系。那么，你可能会问了，机器在训练过程中一定要有标签吗？

其实，机器学习不一定要有标签，具体我们可以分三种情况来看：

训练数据集全部有标签，叫监督学习（supervised learning）；

训练数据集没有标签，叫做无监督学习（unsupervised learning）；

在训练数据集中，有的数据有标签，有的数据没有标签，我们叫做半监督学习（semi-supervised learning）。

目前，监督学习是应用最广泛的主流机器学习算法，所以，我们这个课程的重点会放在监督学习上。当然，对于剩下两种，我也会做个简单的介绍。我们先来着重了解一下什么是监督学习。

什么是监督学习

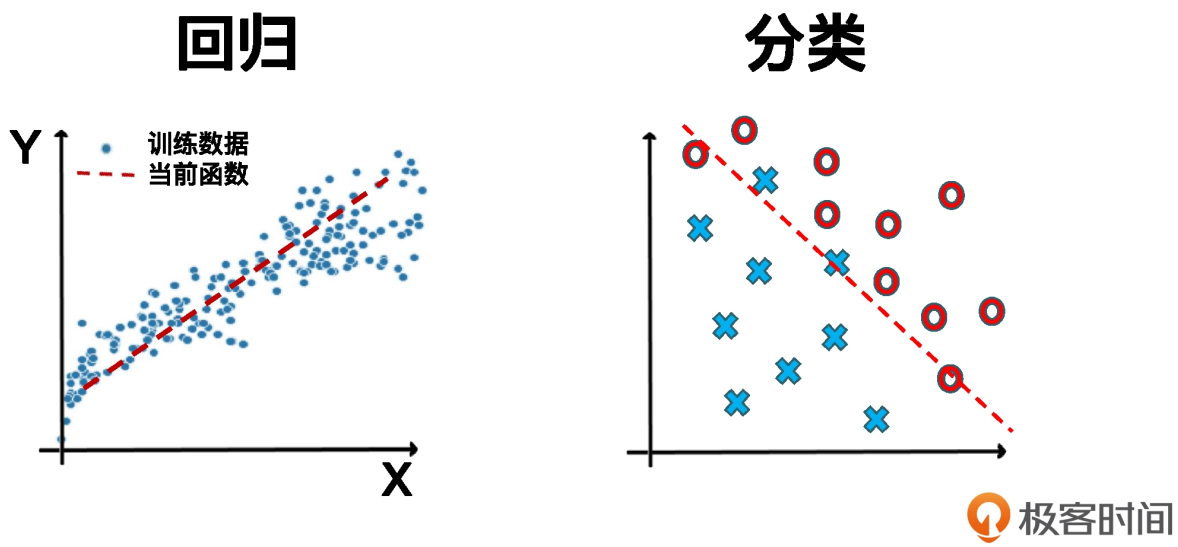
在监督学习中，我们需要重点关注的是监督学习问题的分类。你可能会想问，佳哥，我为什么要知道它的分类？这是因为，明确要解决的问题是机器学习项目的第一步，也是非常重要的一步。如果我们不了解问题的类型，就无法选择合适的算法。

根据标签的特点，监督学习可以被分为两类：回归问题和分类问题。

回归问题的标签是连续数值。比如，如果我们天天给老王的情绪从 1 到 100 打分，那要预测老王今天的情绪，这就是个回归问题。再比如说预测房价，股市，天气情况，这都是回归类型的问题。在我们这个课程里，我特地准备了预测客户的生命周期价值、预测产品转化率等回归项目，你可以在这些项目里学会解决各种回归问题的算法和实战套路。

分类问题的标签是离散性数值。比如，预测老王今天会不会来打牌，这就是个分类问题。而我们平时看到的鉴别高欺诈风险的客户、辅助诊断来访者是否患病、人脸识别等等，这些都属于分类问题的应用。在我们的课程中，我也为你设置了对应的分类实战，包括判断客户是否会流失、判断两款裂变模式哪个更有效等，帮你掌握解决各种分类问题的算法和实战套路。

在这里，我用一张图来给你描绘一下回归和分类的区别：



好，讲完了监督学习，接下来我们简单了解下无监督学习和半监督学习。无监督学习就是为没有标签的数据而建的模型，目前它大多只应用在聚类、降维等有限的场景中，往往是作为数据预处理的一个子步骤显显身手。不过，由于聚类场景相对更加常见，在这门课中，我会用一个“为用户做分组画像”的项目，带你掌握无监督学习的应用。

而半监督学习，就是使用大量无标签数据和一部分有标签数据建模。这往往是因为获取数据标签的难度很高。半监督学习的原理、功能和流程与监督学习是很相似的，区别主要在于多了“伪标签的生成”环节，也就是给无标签的数据人工“贴标签”。

由于半监督学习的落地应用还比较有限，我们的课程中就不专门讲它了，如果你对如何生成伪标签有兴趣的话，可以去研读一些论文。

其实，还有很多现实问题既没法用监督学习来解决，也没法用无监督学习和半监督学习来解决。比如说你要设计一个机器人来陪你玩牌，怎么办？这个时候，就需要强化学习登场了。

什么是强化学习

强化学习研究的目标是，智能体（agent）如何基于环境而做出行动反应，以取得最大化的累积奖励。这里的“智能体”，其实我们可以把它理解成一种机器学习模型。

强化学习和监督学习的差异在于：监督学习是从数据中学习，而强化学习是从环境给它的奖惩中学习。

强化学习智能体在调整策略的时候需要思路比较长远，它不一定每次都明确地选择最优动作，而是要在探索（未知领域）和利用（当前知识）之间找到平衡。它反复试错、不断收集反馈，收集可供自己学习的信号，每经过一个训练周期，都变得比原来强一点，经过亿万次的训练能变得非常强大。

强化学习对数学基础的要求是比较高的，考虑到我们这是面向初学者的机器学习课程，所以这个课程中并没有涉及强化学习的内容。你如果有兴趣，我推荐你阅读一下有关强化学习的专业书籍，比如 Richard Sutton 的经典教材《Reinforcement Learning：An Introduction》，中译名是《强化学习》第 2 版。

从监督学习，到无监督学习，到半监督学习再到强化学习，这就形成了一个完整的闭环。所有机器学习领域中的问题和算法都可以归入其中的某个类别。这里你可能又想问，佳哥，机器学习的分类完全穷尽了吗？我常常听说的深度学习怎么没有在这个分类里呢？别着急，下面我给你解释一下。

什么是深度学习

其实，深度学习是一种使用深层神经网络算法的机器学习模型，也就是一种算法。这个算法可以应用在监督学习、半监督学习和无监督学习里，也可以应用在强化学习中。

虽说深度学习中用的算法叫神经网络算法，但是这个“神经网络”（Artificial Neural Network， ANN）和人脑中的神经网络没啥大的关联，它是数据结构和算法形成的机器学习模型。

我们知道，长期以来，图形图像、自然语言和文本的处理是计算机行业的难题，因为这类信息的数据集，并不是结构化的，需要人工根据信息的类型来选择特征进行提取，这样对于特征的提取是有限的，就拿图像来说，只能提取出一些简单的滤波器。

而深层神经网络的厉害之处在于，它能对非结构的数据集进行自动的复杂特征提取，完全不需要人工干预。也就是说，深度学习让这个曾经的“难题”一下子变得非常容易。

因此，针对深度学习，在后面的课程中，我会带你一起操练三个项目：

使用深层神经网络 DNN 预测用户是否会流失；

使用卷积神经网络 CNN 为鲜花图片分类；

使用循环神经网络 RNN 预测 App 用户的转化率。

学完这三个项目，面对新的类似场景，你就完全能够选择合适的模型，并搭建起属于自己的深层神经网络了。

总而言之，机器学习正在突破我们对“计算机能做什么”这个问题的认知极限。因为机器学习的存在，计算机能做的事情越来越多，以后还会更多。

总结一下

到这里，我们这节课就结束了，不知道你有没有这样一种感受，有些东西乍一听挺唬人的，学了之后发现“嗨，也没啥”。机器学习和深度学习的落地应用，应该可以归为这一类。

在这节课中，我力图给你讲清楚机器学习最基础的一些概念和应用场景。如果你想用一句话给你的朋友说清机器学习是什么，那么请你告诉他：**机器学习是一种从数据生成规则、发现模型，来帮助我们预测、判断、分组和解决问题的技术。**

机器和传统程序最大的不同就是，机器学习不是程序员直接编写函数的技术，是让机器通过“训练”得出函数。而我们做机器学习项目，就是要选定一个算法，然后用数据训练机器，找到一族函数中最适合的那一个，形成模型。

具体来看，机器学习分为四大类，分别是监督学习、无监督学习和半监督学习和强化学习。其中，监督学习是我们课程的一个重点。它能解决的两类问题：**回归和分类**。所以，我们在开始一个项目时，一定要首先明确我们要解决的问题属于哪种类型，这对模型的选择十分重要。

至于深度学习，我们说它是一种使用深层神经网络的模型，可以应用于上述四类机器学习中。深度学习擅长处理非结构化的输入，在视觉处理和自然语言处理方面都很厉害。

最后，我把机器学习中的各种算法做了分类，放到了下图中。这些算法你现在不需要去记忆，以后在项目里用到的时候，我们会解释它们的。

作为过来人，我还想嘱咐一句，我非常理解初学者学机器学习，最怕的就是信息过载，新名词太多，一下子理解不了。其实你不必害怕，我会在后续的课程中，通过实际的项目实战，一步一步手把手，把机器学习中的重要概念和工具陆续教给你，让你在实操过程中把它们搞懂、弄透。

思考题

这节课就到这里了，最后，我想给你留三个思考题，假如你有一个网店：

1. 你有用户的历史订单数据，你想要根据这些数据给用户分组，这是什么类型的机器学习问题？
2. 你有用户的人口统计数据，比如年龄、性别、城市和年收入等等，并且你已经把这些用户已经分成了高、中、低三个价值组。这时拉新团队给了你一批新用户信息，想要判断新用户会属于哪一组，这是什么类型的机器学习问题？
3. 你有用户的人口统计信息和购买历史记录，你想要预测新注册的用户的预期消费总额，这是什么类型的机器学习问题？

提示：先思考是有监督还是无监督学习，如果你认为是监督学习，进一步想想是回归问题、还是分类问题。

欢迎在留言区与我分享你对机器学习的类型的理解，也欢迎你和朋友一起讨论这节课的问题。

极客时间

本讲重点

1. 机器学习是一种从数据中生产函数，而不是程序员直接编写函数的技术。
2. 做机器学习项目，就是明确要解决的问题，再选定一个算法，然后对模型用数据对机器进行训练，找到一族函数中最适合的那一个，形成模型。
3. 机器学习可以分为四大类，分别是监督学习、无监督学习、半监督学习和强化学习。
4. 深度学习是一种使用深层神经网络的算法的模型，它擅长处理非结构化的输入，在视觉处理和自然语言处理方面很厉害。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

赞 6 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 开篇词｜开发者为什么要从实战出发学机器学习？

下一篇 02｜工具准备：安装并使用Jupyter Notebook

精选留言 (13)

写留言

！null 置顶

2021-08-30

买了《数学基础课》《人工智能基础课》《机器学习》《NLP》这些专栏，数学学的很吃力，人工智能基本看不懂，可以考虑从实现学起了

展开

作者回复: 加油，咱们一起动手从做项目中学！

1 5

Yadira 置顶
2021-09-01

聚类、分类、回归
分类和回归的区别：
1、分类的标签是【离散值】；回归的标签的【连续值】；
2、分类实际是找到一个决策面，达到分类的目的；回归实际是找到最优拟合；
3、分类和回归模型的评估方法不同；

展开

作者回复: 精辟，值得置顶。分享给大家。

3

崆峒道 置顶
2021-08-31

这个课的更新频率是什么呢？

展开

编辑回复: 每周一三五，不见不散

1

hk 置顶
2021-08-31

老师666，将难的知识点讲的通俗易懂

展开

作者回复: 666，这是佳哥的强项。

1

有铭 置顶
2021-09-01

预测老王今天会不会来打牌，这就是个分类问题

======

今天突发奇想：如果我每天都预测一下老王来不来打牌。并且这个预测会用历史数据作为自变量，那这个问题是否会从分类问题变成一个回归问题？

展开

作者回复: 不会的。分类还是回归，看的是标签，不是特征。每天都预测一下老王来不来打牌，你的数据集变成了一个时间序列数据集，其中有一个特征字段是日期。但是标签仍然是“是否打牌”。这仍然是二分类。

有铭 置顶

2021-08-31

虽说深度学习中用的算法叫神经网络算法，但是这个“神经网络”（Artificial Neural Network， ANN）和人脑中的神经网络没啥大的关联，它是数据结构和算法形成的机器学习模型

======

我对这句话有疑问，我在别处看到的文献明确的说深度学习的神经网络就是来源于人脑...

展开

作者回复: 谢谢你对此问题提出探讨。是的，很多权威的文献都曾经指出深度学习神经网络结构是来源于对人脑神经理解的启发，包括吴恩达的机器学习课程也这么认为。也有些经典书籍，包括Keras之父的《Python深度学习》一书中指出，二者关联并不大，他的意思可能是说，对于我们初学者来说，不要把神经网络的结构想象的过于复杂，其实每一个深度学习神经元，就是一个基本的逻辑回归模型。

dog_brother 置顶

2021-08-31

一，机器学习的概念：

机器学习是一种从大量的数据中生成规则、发现模型，来帮助我们预测、判断、分组和解决问题的技术。

传统程序是程序员来定义函数，而在机器学习中是机器训练出函数。机器学习的本质特征，就是从数据中发现规则。...

展开

作者回复: 总结的真到位。加油！

小火车不喜欢阴天 置顶

2021-08-30

所以机器学习是弱人工智能吗？

展开

作者回复: 一般认为，强人工智能离我们还很遥远，所以，我的观点是机器学习的确属于弱人工智能。不过，他的实用性可一点都不弱，用好了，能帮我们解决非常多的具体问题。

头发乱了

2021-08-30

1.无监督学习 聚类
2.监督学习 分类
3.监督学习 回归

展开

作者回复: 很正确的答案。手动点赞。

5

keke

2021-08-30

还有迁移学习、集成学习、联邦学习等等，这些“学习们”的关系是啥？

作者回复: 好问题，机器学习新类型很多，分别涉及不同的研究方向。
迁移学习 - 研究如何把机器习得的知识迁移到新模型上？
集成学习 - 我们会有一讲专门讲解。
联邦学习 - 研究如何在进行分布式的机器学习数据共享的同时保护数据隐私等。

还有很多很多的子类别，如图神经网络，自动化机器学习等等。我们将来找一讲专门谈谈这些新领域。

3

！null

2021-08-30

来实践的了，买了再说。

展开

作者回复: 学起来！

2

青松

2021-09-01

1.无监督学习 聚类
2.监督学习 分类
3.监督学习 回归

展开

作者回复: 完全正确！！

1

AIK

2021-08-30

分类，分类，回归

展开

作者回复: 后面两个正确。第一个问题的答案要在思考思考有没有更佳的答案，从题目上看，只有历史订单数据，尚没有已知的分组的信息。

1 1